Summary

By leveraging the Hierarchical Linear Model(HLM), this article aims to examine the effects brought by multi-level influential factors on the housing prices in the Haidian District, Beijing. The regression results indicate that housing prices are negatively correlated with the housing attributes such as living area and year built, meaning people prefer relatively small and new houses. More interestingly, housing buyers are willing to pay high prices on houses with more bedrooms, south facing, first floor and refined decoration.

Introduction

What factors are influencing people's preferences on purchasing houses have been an arresting question for researchers in the fields such as economics, sociology and public policy. In this paper I tried to identify the potential factors which exert impacts on people' valuation on houses that are measured by price per square meter, and also quantify the effects brought by those factors. Furthermore, from the perspective of data structure, the variables such as price per square meter, living area, year built, floor plan, etc. are at transaction level. Meanwhile, there are also grouping variables such as community name, year, month and weekdays. Therefore, leveraging a hierarchical linear model would be an appropriate tool to approach the question. More specifically, I used the price per square meter as the response variable, and utilized continuous variables like living area and year built, and categorical variables such as floor plan, facing direction, floor level and decoration as the predictors. In addition, in order to measure across group variations, I also controlled for grouping variables such as community name, year and month. This paper is organized as follows. This section introduces the question I want to examine and variables we use. The data section illustrates how the dataset is collected and optimized, and also the EDA. In the model section, I try to explain how I select the models and interpret the results. Finally, preliminary conclusions and also limits are about to be discussed in the last section.

Data

Data Pre-processing

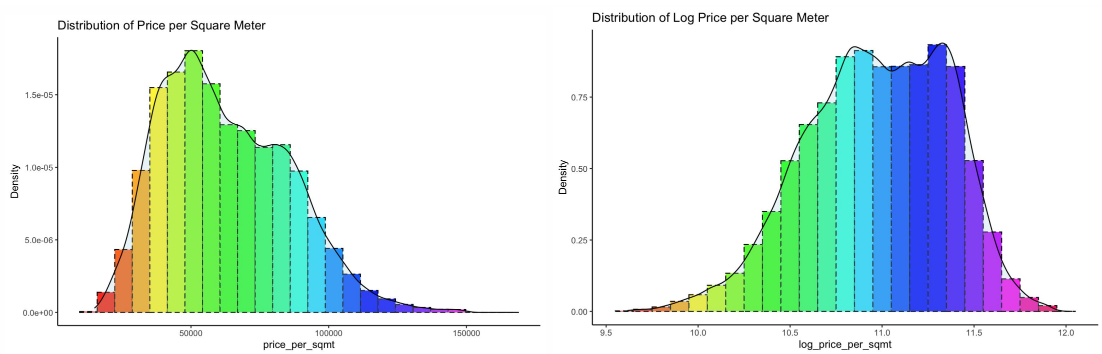
At the beginning, I scraped housing data from the web server of the largest realtor which accounts for over 60% market share of housing transactions in Beijing, and collected 72435 transaction records of Haidian District for the past 8 years, including variables such as price per square meter, community name, livable area, year built, floor plan, facing direction, floor level, decoration, year and month.

Furthermore, I dropped the observations with very low values in terms of price per square meter which could be wrong entries by operators. More importantly, I leveraged the MICE package in R to deal with missing data in the dataset by conducting multiple imputation.

In addition, I loaded the housing dataset, and then categorized the variables such as community name, floor plan, facing direction, floor level, decoration, year and month. Also, I conducted the log transformation for price per square meter.

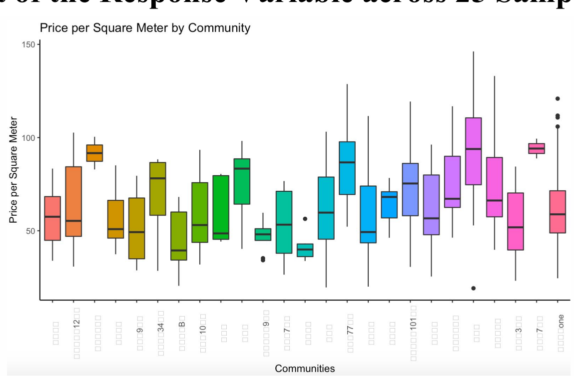
EDA

Figure 1. Distribution of Response Variable before and after transformation



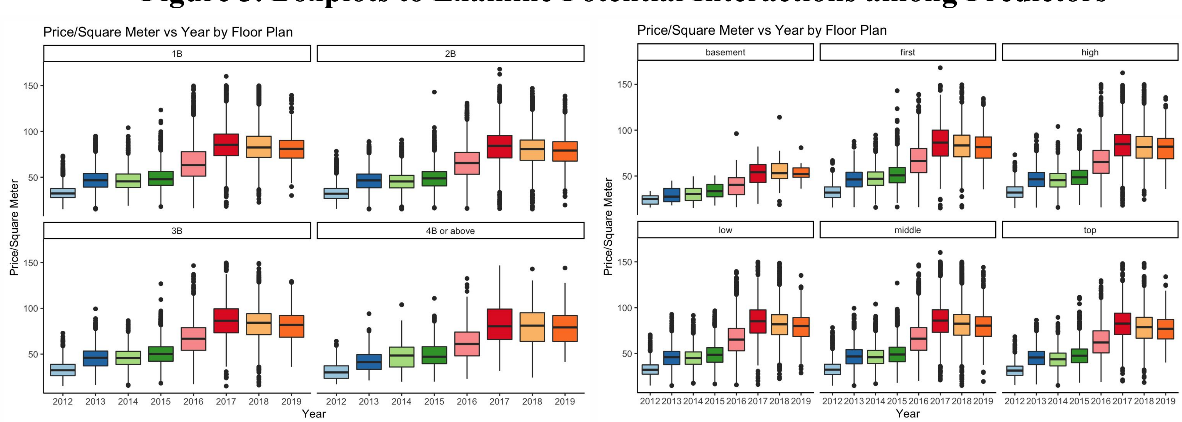
comparing the two plots, I saw that the distribution of response variable is right skewed, and the log transformation counterpart is left skewed with a plateau peak. Thus, I might need to find a better format for response variable later since right now it is not normally distributed.

Figure 2. Boxplot of the Response Variable across 25 Sampled Communities



In order to explore variations across communities, I randomly sampled 25 ones and found that the housing prices varied significantly by community, and there are communities with little data.

Figure 3. Boxplots to Examine Potential Interactions among Predictors



In addition, I also examined potential interactions and found that there may not be interactions between housing attributes such as floor plan and floor level and time-related variables like year.

Model

Frequentist Approach

Based on the grouped data structure discussed in introduction and small data points within many communities revealed in EDA, I was convinced that a hierarchical linear model is suitable in this situation, and can make unbiased estimates for the fixed effects and overall variance parameters.

I began our modeling process with a baseline model by using price per square meter as the response variable, and having predictors including living area, year built, floor plan, facing direction, floor level and decoration at housing transactions level(lower level). In addition, in order to consider variations across groups, I was also controlling for group-level(higher level) predictors such as community name, year, month and weekday. I then leveraged ANOVA tests to consider predictors and interactions individually, and removed the predictor "weekday". After the process, I arrived at my first hierarchical model below.

$$\begin{aligned} \textit{price per square meter}_{ijkh} = {} & (\beta_0 + \gamma_{0j} + \eta_{0k} + \text{ø}_{0h}) + \beta_1 \textit{living area}_{ijkh} + \beta_2 \textit{year built}_{ijkh} + \beta_3 \textit{floor plan}_{ijkh} \\ & + \beta_4 \textit{facing direction}_{ijkh} + \beta_5 \textit{floor level}_{ijkh} + \beta_6 \textit{decoration}_{ijkh} + \varepsilon_{ijkh} \end{aligned}$$

$$\gamma_{0j} \sim N(0,\ \tau^2_{\gamma(0)}) \quad \eta_{0k} \sim N(0,\ \tau^2_{\eta(0)}) \quad \text{ø}_{0h} \sim N(0,\ \tau^2_{\text{ø}(0)}) \quad \varepsilon_{ijh} \sim N(0, \sigma^2)$$

$$i = 1, \dots, n_{jkh};\ j = 1, \dots, J;\ k = 1, \dots, K;\ h = 1, \dots, H.$$

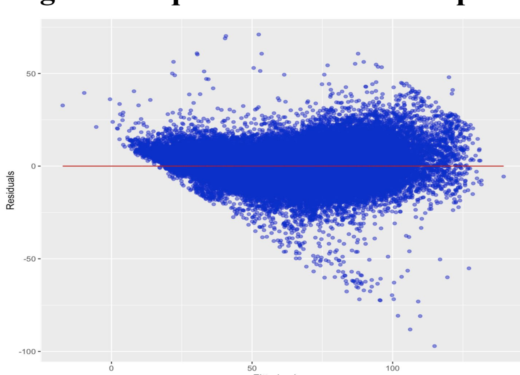
After creating the baseline model, I checked the model assumptions. Specifically, I found the independence and equal variance assumption violated(lower left plot). Thus, I tried Boxcox Transformation for the response variable, and got improvement for the result in assumption validation, but made the results hard to interpret. Furthermore, I chose log transformation for the response, and achieved similar results for model validation(lower right plot) like its peer in the Boxcox one but made the interpretation easier. Finally, I created my improved model below.

$$\begin{aligned} \log(\textit{price per square meter}_{ijkh}) = {} & (\beta_0 + \gamma_{0j} + \eta_{0k} + \text{ø}_{0h}) + \beta_1 \textit{living area}_{ijkh} + \beta_2 \textit{year built}_{ijkh} + \beta_3 \textit{floor plan}_{ijkh} \\ & + \beta_4 \textit{facing direction}_{ijkh} + \beta_5 \textit{floor level}_{ijkh} + \beta_6 \textit{decoration}_{ijkh} + \varepsilon_{ijkh} \end{aligned}$$

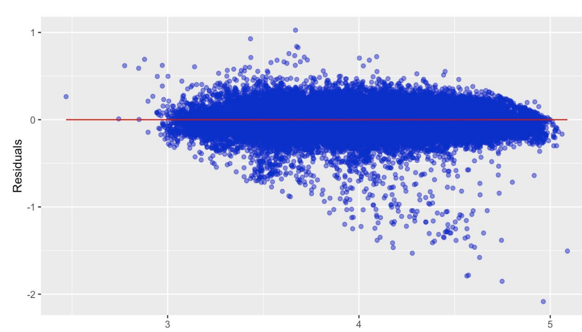
$$\gamma_{0j} \sim N(0,\ \tau^2_{\gamma(0)}) \quad \eta_{0k} \sim N(0,\ \tau^2_{\eta(0)}) \quad \text{ø}_{0h} \sim N(0,\ \tau^2_{\text{ø}(0)}) \quad \varepsilon_{ijh} \sim N(0, \sigma^2)$$

$$i = 1, \dots, n_{jkh};\ j = 1, \dots, J;\ k = 1, \dots, K; h = 1, \dots, H.$$

Figure 4. Equal Variance Assumption Checking before and after Transformation



Residuals vs Fitted values



Residuals vs Fitted values

Bayesian Approach

After utilizing the lme4 package in R for building the previous model, I was still concerned about its validity since the frequentist approach used by the lme4 package does not fully account for uncertainty in the estimated variance parameters. Therefore, I tried to leverage the Bayesian method to achieve a better estimation below.

Table 1. Bayesian Inference for the Transaction-Level Effects

	effect	component	term	estimate	std.error	conf.low	conf.high
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	fixed	cond	(Intercept)	41.1	7.87	25.3	56.9
2	fixed	cond	living_area	-0.113	0.00198	-0.117	-0.109
3	fixed	cond	year_built	-0.0760	0.00895	-0.0927	-0.0572
4	fixed	cond	floor_plan2B	0.402	0.0970	0.213	0.587
5	fixed	cond	floor_plan3B	2.22	0.141	1.95	2.49
6	fixed	cond	floor_plan4Borabove	4.72	0.308	4.10	5.30
7	fixed	cond	facing_directionNorth	-0.255	0.144	-0.523	0.0359
8	fixed	cond	facing_directionSouth	3.56	0.102	3.35	3.75
9	fixed	cond	floor_levelfirst	28.7	0.470	27.8	29.6
10	fixed	cond	floor_levelhigh	28.2	0.463	27.3	29.1
11	fixed	cond	floor_levellow	28.0	0.464	27.0	28.9
12	fixed	cond	floor_levelmiddle	28.6	0.461	27.6	29.4
13	fixed	cond	floor_leveltop	26.0	0.467	25.0	26.9
14	fixed	cond	decorationsimple	-0.867	0.0709	-0.999	-0.718

Unfortunately, after 18 hours of torturous processing, I found that some parameters underscored by the red color above may not be reliable since the report warning showed that the default iteration numbers and the maximum treedepth were not enough for running the model. Since I don't have another 18 hours to rerun the model with modified parameters, I will go back to the log transformation model since it is at least able to make unbiased estimates for the fixed effects.

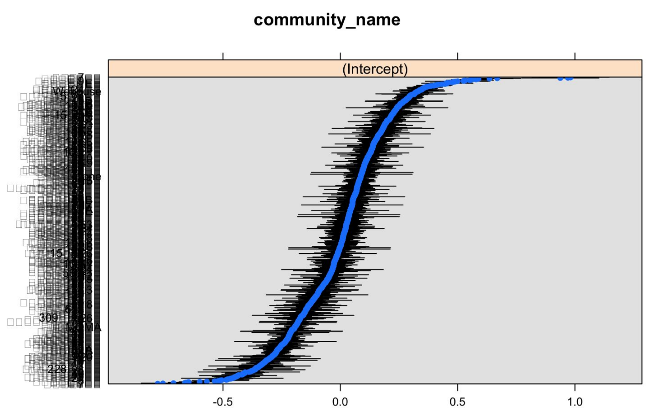
Results

Table 2. Regression Results of Frequentist Inference

	log_price_per_sqmt		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.64	3.39 – 3.89	**<0.001**
living_area	-0.00	-0.00 – -0.00	**<0.001**
year_built	-0.00	-0.00 – -0.00	**<0.001**
floor_plan [2B]	0.01	0.01 – 0.02	**<0.001**
floor_plan [3B]	0.04	0.04 – 0.05	**<0.001**
floor_plan [4B or above]	0.08	0.08 – 0.09	**<0.001**
facing_direction [North]	-0.01	-0.01 – -0.00	**0.011**
facing_direction [South]	0.06	0.05 – 0.06	**<0.001**
floor_level [first]	0.54	0.53 – 0.56	**<0.001**
floor_level [high]	0.54	0.52 – 0.55	**<0.001**
floor_level [low]	0.53	0.52 – 0.55	**<0.001**
floor_level [middle]	0.54	0.53 – 0.56	**<0.001**
floor_level [top]	0.50	0.48 – 0.51	**<0.001**
decoration [simple]	-0.01	-0.02 – -0.01	**<0.001**
Random Effects			
σ^2	0.02		
$\tau_{00\ community\_name}$	0.05		
$\tau_{00\ month}$	0.00		
$\tau_{00\ year}$	0.13		
ICC	0.91		
$N_{community\_name}$	1126		
N_{year}	8		
N_{month}	12		
Observations	72166		
Marginal R^2 / Conditional R^2	0.027 / 0.917		

According to the above regression table, the baseline averaged housing price per square meter was exp(3.64), which was 38.09 thousand Yuan in RMB for a transaction in Haidian District with one bedroom, east/west facing, basement level and refined decoration over the past 8 years. Furthermore, I found that the living area and year built were associated with significant negative effects on the response variable. Specifically, Living area has a coefficient of -0.00198, resulting in an exponentiated point estimate of 0.998 which means one square meter increase in living area would be correlated with 0.2% reduction in housing price per square meter when holding other predictors constant. Meanwhile, the year built has a coefficient of -0.000782 meaning one year older of a house would reduce price per square meter by 0.1%. More interestingly, I found that the more bedrooms a house had, the higher its price per square meter was. For example, the price per square meter of houses with 3 bedrooms would be 4.5% higher than its peers with one bedroom. Moreover, buyers valued the south-facing and middle level houses most, and likely spent 6.0% and 72.1% more than the baseline peers with regard to the housing prices respectively. In addition, people would be willing to pay 1.5% higher for refined houses compared to their peers with simple decoration in terms of price per square meter.

Figure 5. Dotplot for Variations across Communities



Furthermore, I also found that the values of the price per square meter of housing transactions varied across communities. Looking at the dot-plot above, I noticed that many communities on two ends of the plot that do not contain 0 at the 95% confidence interval. Especially, some communities at the upper right are very far from their peers in terms of the mean value, confirming that the housing price per square meter does vary across the communities.

Conclusion

Using hierarchical linear models, this paper tried to examine the impact of influential factors such as housing attributes and time periods on housing prices in the Haidian District, Beijing. The results revealed significant preferences of people on choosing housing. However, the models used still did not perfectly satisfy the independent and equal variance assumption which would weaken credibility for the inference results. In addition, there is a silver lining that I can utilize the Bayesian model to make more reliable inference than previous models did in further study.

Appendix

Tables

Table 1. Regression Results by Using REML

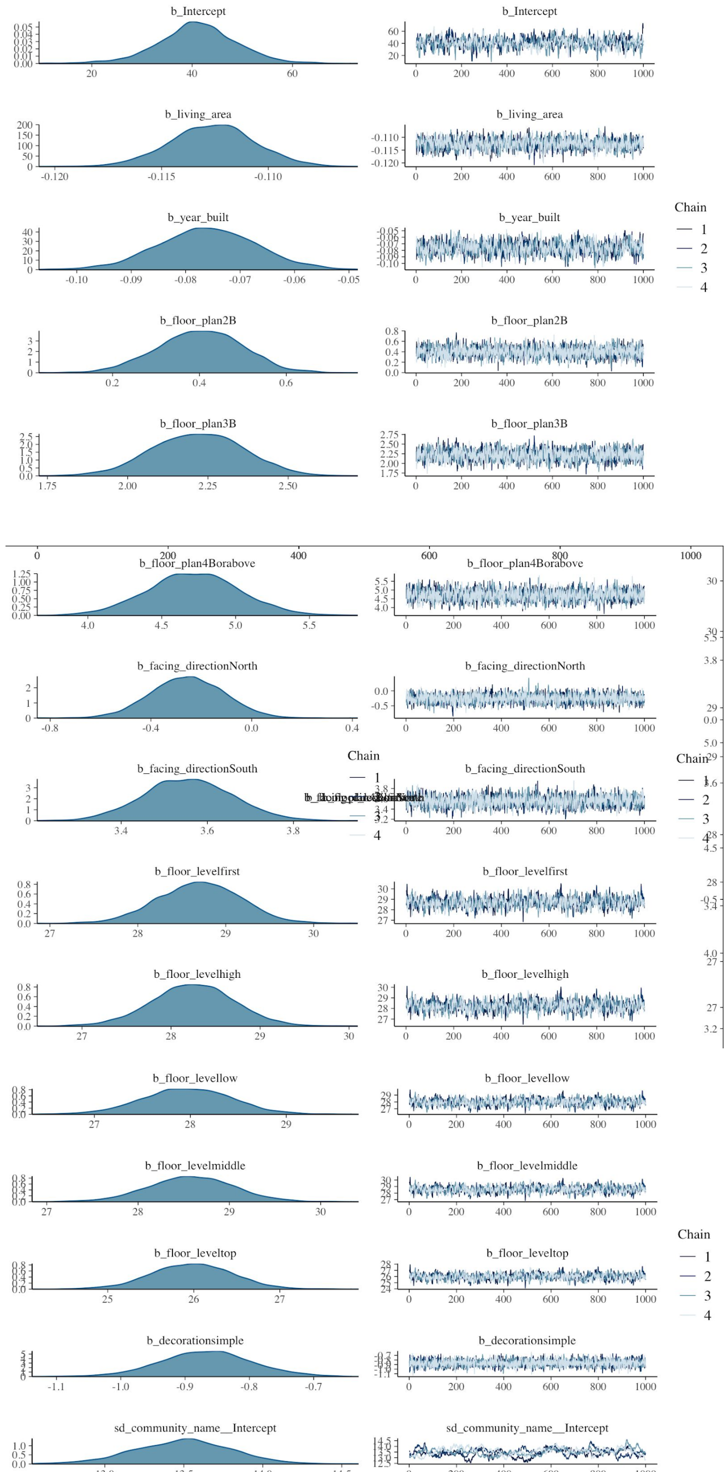
```
Linear mixed model fit by REML ['lmerMod']
Formula: log_price_per_sqmt ~ living_area + year_built + floor_plan +
    facing_direction + floor_level + decoration + (1 | community_name) +
    (1 | year) + (1 | month)
   Data: housing

REML criterion at convergence: -86490.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-15.7518  -0.5022   0.0365   0.5495   8.7551

Random effects:
 Groups         Name        Variance Std.Dev.
 community_name (Intercept) 0.047682 0.21836
 month          (Intercept) 0.001909 0.04369
 year           (Intercept) 0.126373 0.35549
 Residual                   0.016408 0.12809
Number of obs: 72166, groups:  community_name, 1126; month, 12; year, 8

Fixed effects:
                        Estimate Std. Error t value
(Intercept)            3.640e+00  1.268e-01  28.714
living_area           -1.980e-03  3.106e-05 -63.763
year_built            -7.823e-04  1.422e-04  -5.500
floor_plan2B           1.406e-02  1.498e-03   9.386
floor_plan3B           4.371e-02  2.173e-03  20.114
floor_plan4B or above  8.451e-02  4.790e-03  17.642
facing_directionNorth -5.589e-03  2.200e-03  -2.540
facing_directionSouth  5.784e-02  1.611e-03  35.899
floor_levelfirst       5.418e-01  7.508e-03  72.166
floor_levelhigh        5.370e-01  7.434e-03  72.229
floor_levellow         5.321e-01  7.442e-03  71.494
floor_levelmiddle      5.430e-01  7.400e-03  73.377
floor_leveltop         4.989e-01  7.475e-03  66.739
decorationsimple      -1.480e-02  1.091e-03 -13.570
```

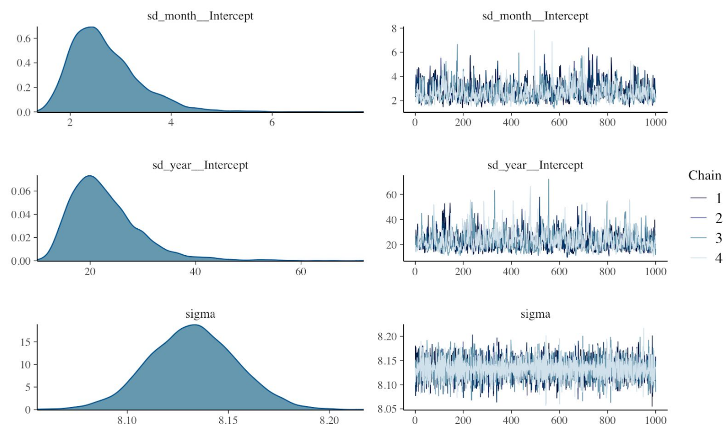
Table 2. Regression Results by Using Bayesian Method

```
Group-Level Effects:
~community_name (Number of levels: 1126)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)    13.50      0.31    12.88    14.08 1.08       70      180

~month (Number of levels: 12)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     2.69      0.68     1.73     4.26 1.00      707     1370

~year (Number of levels: 8)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)    22.86      6.73    13.67    39.81 1.01      842     1320

Population-Level Effects:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                41.07      7.87    24.91    56.51 1.00      606     1201
living_area              -0.11      0.00    -0.12    -0.11 1.00     1538     2177
year_built               -0.08      0.01    -0.09    -0.06 1.00      731     1489
floor_plan2B              0.40      0.10     0.21     0.58 1.00     1858     2702
floor_plan3B              2.22      0.14     1.94     2.49 1.00     1719     2443
floor_plan4Borabove       4.72      0.31     4.12     5.32 1.00     1908     2627
facing_directionNorth    -0.25      0.14    -0.54     0.03 1.00     2862     3288
facing_directionSouth     3.56      0.10     3.36     3.76 1.00     2517     3043
floor_levelfirst         28.70      0.47    27.78    29.60 1.00      803     1721
floor_levelhigh          28.25      0.46    27.33    29.16 1.00      791     1609
floor_levellow           27.97      0.46    27.05    28.88 1.00      786     1641
floor_levelmiddle        28.58      0.46    27.67    29.49 1.00      777     1525
floor_leveltop           25.98      0.47    25.04    26.89 1.00      777     1601
decorationsimple         -0.87      0.07    -1.01    -0.72 1.00     4538     2916

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     8.13      0.02     8.09     8.18 1.00    10546     2839

Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

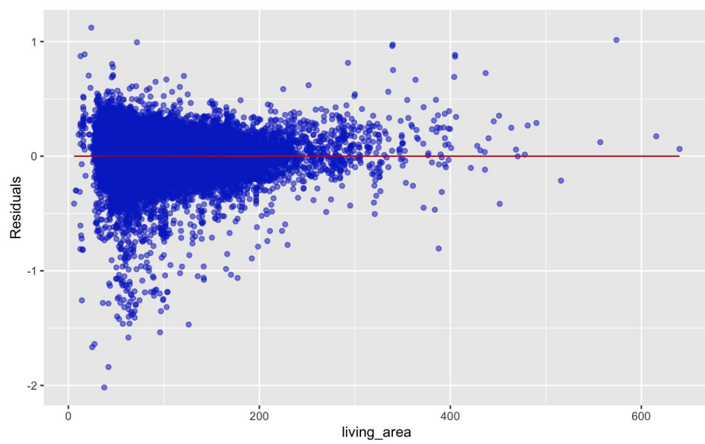
Plots

Plot 1. Plots of Results by Using Bayesian Method



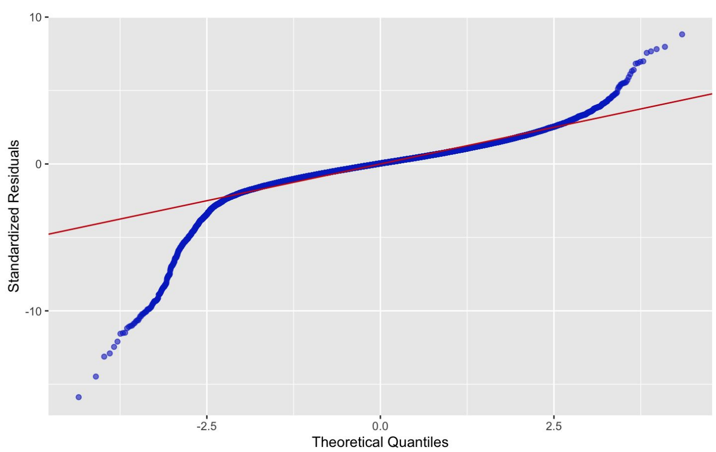


Plot 2. Linear Assumption Checking by Using Living Area



Residuals vs living_area

Plot 3. Normality Assumption Checking



Normal Q-Q

Code

```
###### Clear environment and load libraries
rm(list = ls())
library(ggplot2) #Time to pivot to ggplot2
library(lme4)
library(rstan)
library(brms)
library(broom)
library(sjPlot) #another option for making nice html tables
library(lattice) #for ranef
library(magrittr)
library(dplyr)
library(knitr)
library(xtable)
library(kableExtra)
library(stargazer)
library(rms) #for VIF
library(MASS)
library(arm)
library(pROC)
library(e1071)
library(caret)
require(gridExtra)
library(mice) #for missing data imputation

##################################################
# Load the data
##################################################

housing <- read.csv(file= "/Users/Reinhard/702_Statistical_Modelling_Final_Project/df9.csv",
header=TRUE, sep=",")
summary(housing)
str(housing)

##################################################
# Data Preparation
##################################################
```

```
# Missing Data
##################################################

housing$community_name_f <- factor(housing$community_name)

housing$floor_plan_f <- factor(housing$floor_plan)

housing$facing_direction <- factor(housing$facing_direction)

housing$floor_level <- factor(housing$floor_level)

housing$decoration <- factor(housing$decoration)

housing$year_f <- factor(housing$year)

housing$month_f <- factor(housing$month)

housing$weekday_name_f <- factor(housing$weekday_name)

dim(housing)
str(housing)
summary(housing)

md.pattern(housing)

housing_imp <- mice(housing,m=3,
                 defaultMethod=c("norm","logreg","polyreg","polr"),
                 print=F)

densityplot(housing_imp)

d1 <- complete(housing_imp, 1)
d1

str(d1)
write.csv(d1, file = "d1_imp.csv")

##################################################
# Reload the data
##################################################
```

```
housing <- read.csv(file= "/Users/Reinhard/702_Statistical_Modelling_Final_Project/df14.csv",
header=TRUE, sep=",")
summary(housing)
str(housing)

# Catagorizing Variables
##################################################

housing$community_name <- factor(housing$community_name)

housing$floor_plan <- factor(housing$floor_plan)

housing$facing_direction <- factor(housing$facing_direction)

housing$floor_level <- factor(housing$floor_level)

housing$decoration <- factor(housing$decoration)

housing$year <- factor(housing$year)

housing$month <- factor(housing$month)

housing$weekday_name <- factor(housing$weekday_name)

# Continuous Variable
##################################################
housing$year_built <- 2020 - housing$year_built
str(housing$year_built)

housing$living_area <- housing$livable_area
str(housing$living_area)

housing$price_per_sqmt <- housing$price_per_sqmt/1000

housing$log_price_per_sqmt <- log(housing$price_per_sqmt)

housing$sqrt_price_per_sqmt <- sqrt(housing$price_per_sqmt)

housing$log_living_area <- log(housing$living_area)
```

```r
housing$log_year_built <- log(housing$year_built)

# Mean Centering Continous Variables
####################################################
housing$livable_area_c <- housing$livable_area - mean(housing$livable_area)

housing$year_built_c <- housing$year_built - mean(housing$year_built)


####################################################
# EDA
####################################################

str(housing)

dev.off()

#histogram
#####################################################

ggplot(housing,aes(x=price_per_sqmt)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                 fill=rainbow(25), bins = 25) +
  geom_density(alpha=.25, fill="lightblue") +
  scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Price per Square Meter",y="Density") +
  theme_classic() + theme(legend.position="none")


ggplot(housing,aes(x=log_price_per_sqmt)) +
  geom_histogram(aes(y=..density..),color="black",linetype="dashed",
                 fill=rainbow(25),binwidth = 0.1) +
  geom_density(alpha=.25, fill="lightblue") +
  scale_fill_brewer(palette="Blues") +
  labs(title="Distribution of Log Price per Square Meter",y="Density") +
  theme_classic() + theme(legend.position="none")

#Sampling
#####################################################
```

```
set.seed(1000)
sample_community <- sample(unique(housing$community_name),25,replace=F)

ggplot(housing[is.element(housing$community_name,sample_community),],
    aes(x=community_name, y=price_per_sqmt, fill=community_name)) +
    geom_boxplot() +
    labs(title="Price per Square Meter by Community",
    x="Communities",y="Price per Square Meter") + theme_classic() +
    theme(legend.position="none",axis.text.x = element_text(angle = 90))

ggplot(housing, aes(living_area, y=price_per_sqmt)) +
  geom_point(alpha = .5,colour="blue3") +
  geom_line(y=0, col="green3") +
  geom_smooth(method = "lm", col = "red3") + xlab("Living Area") +
  ylab("Price/Square Meter") +
  #labs(caption="Price/Square Meter vs Living Area") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20)) +
  facet_wrap(~year,ncol = 5)

#Box Plots
#####################################################
ggplot(housing,aes(x=year, y=price_per_sqmt, fill=year)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Paired") +
  labs(title="Price/Square Meter vs Year by Floor Plan",x="Year",y="Price/Square Meter") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ floor_plan)

ggplot(housing,aes(x=year, y=price_per_sqmt, fill=year)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Paired") +
  labs(title="Price/Square Meter vs Year by Floor Plan",x="Year",y="Price/Square Meter") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap( ~ floor_level)

####################################################
# Model fitting
####################################################

###### MLR
```

```
#####################################################

model1 <- lm(price_per_sqmt ~ livable_area, data=housing)
summary(model1)

model2 <- lm(price_per_sqmt ~ livable_area + year_built, data=housing)
summary(model2)

model3 <- lm(price_per_sqmt ~ livable_area + year_built + floor_plan, data=housing)
summary(model3)

model4 <- lm(price_per_sqmt ~ livable_area + year_built + floor_plan + facing_direction,
data=housing)
summary(model4)

model5 <- lm(price_per_sqmt ~ livable_area + year_built + floor_plan + facing_direction +
floor_level, data=housing)
summary(model5)

model6 <- lm(price_per_sqmt ~ livable_area + year_built + floor_plan + facing_direction +
floor_level + decoration, data=housing)
summary(model6)

model61 <- lm(price_per_sqmt ~ livable_area + year_built + floor_plan + facing_direction +
floor_level + decoration + community_name + year + month, data=housing)
summary(model61)

###### Boxcox Transformation
#####################################################

boxcox(model61)

boxcox(model61, lambda = seq(-0.25, 0.75, by = 0.05),  plotit = TRUE)

model61_cox <- lm((price_per_sqmt^0.3 - 1) / 0.3 ~ livable_area + year_built + floor_plan +
facing_direction + floor_level + decoration + community_name + year + month, data=housing)
summary(model61_cox)

res1 <- residuals(model61_cox,"resp")
```

```
pred1 <- predict(model61_cox)

pred_res1 <- data.frame(pred1, res1)
var_plot <- ggplot(pred_res1, aes(pred1, y=res1)) +
  geom_point(alpha = .5,colour="blue3") +
  geom_line(y = 0, col = "red3") +
  geom_smooth(method="loess",col="red3") +
  xlab("Fitted values") +
  ylab("Residuals") +
  labs(caption="Residuals vs Fitted values") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))
var_plot

std_res1 <- (res1 - mean(res1)) / sd(res1)
std_res_df1 <- data.frame(std_res1)
q_plot <- qplot(sample = std_res1, data = std_res_df1, color=I("blue3"), alpha=.5) +
  geom_abline(intercept = 0, slope = 1, col="red3") +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  labs(caption="Normal Q-Q") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20), legend.position = "none")
q_plot

###### Hierarchical Linear Model
####################################################

model7 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + year + month + weekday_name + (1 |
community_name ) )
summary(model7)
tab_model(model7)

model71 <- lmer(data = housing, log_price_per_sqmt ~ log_living_area + year_built +
floor_plan + facing_direction + floor_level + decoration + year + month + weekday_name + (1 |
community_name ) )
summary(model7)
tab_model(model71)
```

```
model8 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1|year) + month + weekday_name + (1 |
community_name ))
summary(model8)
tab_model(model8)

model81 <- lmer(data = housing, log_price_per_sqmt ~ log_living_area + year_built +
floor_plan + facing_direction + floor_level + decoration + (1|year) + month + weekday_name +
(1 | community_name ))
summary(model81)
tab_model(model81)

model9 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1|month))
summary(model9)

model10 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1|weekday_name))
summary(model10)

model11 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1 | community_name) + (1|year))
summary(model11)

###Main Effects
####################################################

str(housing)

model12 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1 | community_name) + (1|year) + (1|month))
summary(model12)
tab_model(model12)

model13 <- lmer(data = housing, price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1 | community_name) + (1|year) + (1|month)+
(1|weekday_name))
summary(model13)
```

```
model14 <- lmer(data = housing, log_price_per_sqmt ~ living_area + year_built + floor_plan +
facing_direction + floor_level + decoration + (1 | community_name) + (1|year) + (1|month))
summary(model14)
tab_model(model14)
dotplot(ranef(model14, condVar = TRUE))

model15 <- lmer(data = housing, log_price_per_sqmt ~ log_living_area + year_built +
floor_plan + facing_direction + floor_level + decoration + (1 | community_name) + (1|year) +
(1|month))
summary(model15)
tab_model(model15)

model16 <- lmer(data = housing, log_price_per_sqmt ~ log_living_area + year_built +
floor_plan + facing_direction + floor_level + decoration + (1 | community_name) + (1|year) +
(1|month) + (1|weekday_name))
summary(model16)
tab_model(model16)

model17 <- lmer(data = housing, log_price_per_sqmt ~ log_living_area + log_year_built +
floor_plan + facing_direction + floor_level + decoration + (1 | community_name) + (1|year) +
(1|month))
summary(model17)
tab_model(model17)

model18 <- lmer(data = housing, log_price_per_sqmt ~ log_living_area + (1 | year_built) +
floor_plan + facing_direction + floor_level + decoration + (1 | community_name) + (1|year) +
(1|month))
summary(model18)
tab_model(model18)

anova(model7, model11)

anova(model11, model12)

anova(model12, model13)

anova(model12, model14)

anova(model14, model15)
```

```
anova(model15, model16)

###### Interactions
################################################

###Livable_area* others
####################
model19 <- lmer(data = housing, price_per_sqmt ~ livable_area_c*(floor_plan +
facing_direction + floor_level + decoration + year_built)  + (1 | community_name) + (1|year) +
(1|month))
summary(model19)
tab_model(model19) #nonsense: floor_plan,interactions

anova(model12, model19)

model20 <- lmer(data = housing, log_price_per_sqmt ~ livable_area_c*(floor_plan +
facing_direction + floor_level + decoration + year_built)  + (1 | community_name) + (1|year) +
(1|month))
summary(model20)
tab_model(model20) #nonsense: floor_plan,interactions

model21 <- lmer(data = housing, price_per_sqmt ~ livable_area_c*(year_built +
facing_direction + floor_level + decoration) + floor_plan + (1 | community_name) + (1|year) +
(1|month))
summary(model21)
tab_model(model21)  # small interactions, year_built sign change

anova(model12, model21)

model22 <- lmer(data = housing, price_per_sqmt ~ livable_area_c*( facing_direction +
year_built ) + floor_level + decoration + floor_plan  + (1 | community_name) + (1|year) +
(1|month))
summary(model22)
tab_model(model22) # nonsense or small interactions

model23 <- lmer(data = housing, log_price_per_sqmt ~ livable_area_c*( facing_direction +
year_built ) + floor_level + decoration + floor_plan  + (1 | community_name) + (1|year) +
(1|month))
summary(model23)
tab_model(model23) # nonsense or small interactions
```

```
anova(model12, model23)

str(housing)

model24 <- lmer(data = housing, price_per_sqmt ~ livable_area_c*( decoration +
facing_direction ) + floor_level + year_built + floor_plan  + (1 | community_name) + (1|year) +
(1|month))
summary(model24)
tab_model(model24) # small interactions

anova(model12, model24)

model25 <- lmer(data = housing, log_price_per_sqmt ~ livable_area_c*( decoration +
facing_direction ) + floor_level + year_built + floor_plan  + (1 | community_name) + (1|year) +
(1|month))
summary(model25)
tab_model(model25) # small interactions


model26 <- lmer(data = housing, price_per_sqmt ~ livable_area_c*( decoration +
facing_direction ) + floor_level*floor_plan + year_built + (1 | community_name) + (1|year) +
(1|month))
summary(model26)
tab_model(model26) # small interactions

model27 <- lmer(data = housing, log_price_per_sqmt ~ livable_area_c*( decoration +
facing_direction ) + floor_level*floor_plan + year_built + (1 | community_name) + (1|year) +
(1|month))
summary(model27)
tab_model(model27) # small interactions

# Dotplot
####################################################
dotplot(ranef(model14, condVar = TRUE))

###Bayesian Approach
####################################################

#Model conducted
```

```
#################
model_bayes <- brm(price_per_sqmt ~ living_area + year_built + floor_plan +
                   facing_direction + floor_level + decoration + (1 | community_name) +
                   (1|year) + (1|month), data=housing)
summary(model_bayes)
head(predict(model_bayes))
dim(posterior_samples(model_bayes))
#head(posterior_samples(model_bayes))

library(broom.mixed)
tidy(model_bayes)
tidy(model_bayes, parameters = "^sd_", conf.int = FALSE)
tidy(model_bayes, effects = "fixed", conf.method="HPDinterval")

#visualize results: plot estimated OR's with 2 SD error bars ea. side
plot_model(model_bayes,show.values = TRUE)
plot_model(model_bayes,type="pred")
p <- plot_model(model_bayes,type="re",show.values = TRUE,ri.nr = c(1,2))

#Model for next run
#################
model_bayes1 <- brm(log_price_per_sqmt ~ living_area + year_built + floor_plan +
                    facing_direction + floor_level + decoration + (1 | community_name) +
                    (1|year) + (1|month), data=housing,iter = 4000, chains = 4,
                    control = list(adapt_delta = 0.995, max_treedepth = 20))
summary(model_bayes1)
head(predict(model_bayes1))
dim(posterior_samples(model_bayes1))

###############################################
# Model Assumptions Checking
###############################################

res <- residuals(model14,"resp")

pred <- predict(model14)

#Linearity
###############################################
ggplot(housing, aes(living_area, y=res)) +
```

```
  geom_point(alpha = .5,colour="blue3") +
  geom_line(y=0, col="red3") +
    xlab("livable_area") +
    ylab("Residuals") +
  labs(caption="Residuals vs log_livable_area") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))


ggplot(housing, aes(year_built, y=res)) +
  geom_point(alpha = .5,colour="blue3") +
  geom_line(y=0, col="red3") +
  xlab("livable_area") +
  ylab("Residuals") +
  labs(caption="Residuals vs year built") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))

# Independence and Equality of Variance
####################################################
pred_res <- data.frame(pred, res)
var_plot <- ggplot(pred_res, aes(pred, y=res)) +
  geom_point(alpha = .5,colour="blue3") +
  geom_line(y = 0, col = "red3") +
  geom_smooth(method="loess",col="red3") +
  xlab("Fitted values") +
  ylab("Residuals") +
  labs(caption="Residuals vs Fitted values") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))
var_plot

# Normality
####################################################
std_res <- (res - mean(res)) / sd(res)
std_res_df <- data.frame(std_res)
q_plot <- qplot(sample = std_res, data = std_res_df, color=I("blue3"), alpha=.5) +
  geom_abline(intercept = 0, slope = 1, col="red3") +
  xlab("Theoretical Quantiles") +
  ylab("Standardized Residuals") +
  labs(caption="Normal Q-Q") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20), legend.position = "none")
q_plot
```

```
# Check outlier with high influencial
##################################################
cook <- cooks.distance(model12)
lev <- hatvalues(model12)
cookd <- data.frame(lev, std_res, cook)
ggplot(cookd, aes(lev, std_res)) +
  geom_point(aes(size=cook), col="blue3", alpha=.5) +
  geom_smooth(method="loess", col="red3") +
  xlab("Leverage") +
  ylab("Standardized Residuals") +
  labs(caption="Standardized Residuals vs Leverage") +
  theme(plot.caption = element_text(hjust = 0.5, size = 20))
```